Adventures in the jungle of high-order optimization

Philippe Toint (with Coralia Cartis and Nick Gould)

Namur Center for Complex Systems (naXys), University of Namur, Belgium

(philippe.toint@unamur.be)

Toulouse, April 2017

The problem

We consider the unconstrained nonlinear programming problem:

$$\text{minimize} \quad f(x)$$

for $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ smooth.

Important special case: the nonlinear least-squares problem

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\|F(x)\|^2$$

for $x \in \mathbb{R}^n$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ smooth.

Using degree 2 models

Note the following: if f has a globally Lipschitz continuous Hessian H with constant $2L$, define

$$T_{f,2}(x,s) \stackrel{\text{def}}{=} f(x) + \langle g(x), s\rangle + \frac{1}{2}\langle s, H(x)s\rangle$$

Taylor, Cauchy-Schwarz and Lipschitz then imply

$$\begin{aligned} f(x+s) &= T_{f,2}(x,s) \\ &\quad + \int_0^1 (1-\alpha)\langle s, \Big(H(x+\alpha s) - H(x)\Big)s\rangle \, d\alpha \\ &\leq T_{f,2}(x,s) + \tfrac{1}{3}L\|s\|_2^3 \stackrel{\text{def}}{=} m(s) \end{aligned}$$

$\Longrightarrow$ reducing m from $s = 0$ improves f since $m(0) = f(x)$
$\Longrightarrow$ cubically regularized quadratic model

Using high-degree models

Note the following: if f has a globally Lipschitz continuous p-th derivative tensor $\nabla_x^p f$ with constant $(p-1)!L$, define

$$T_{f,p}(x,s) \stackrel{\text{def}}{=} f(x) + \sum_{j=1}^{p} \frac{1}{j!} \nabla_x^j f(x)[s]^j$$

Taylor, Cauchy-Schwarz and Lipschitz then imply

$$\begin{aligned} f(x+s) &= T_{f,p}(x,s) \\ &\quad + \tfrac{1}{(p-1)!} \int_0^1 (1-\alpha)^{p-1} \Big(\nabla_x^p(x+\alpha s) - \nabla_x^p(x) \Big)[s]^p \, d\alpha \\ &\leq T_{f,p}(x,s) + \tfrac{1}{p} L \|s\|_2^{p+1} \stackrel{\text{def}}{=} m(s) \end{aligned}$$

$\Longrightarrow$ reducing m from $s=0$ improves f since $m(0) = f(x)$
$\Longrightarrow$ $(p+1)$th-power regularized degree-p model

Approximate model minimization

Lipschitz constant L unknown $\Rightarrow$ replace by adaptive parameter σ_k in the model :

$$m(s) \stackrel{\text{def}}{=} T_{f,p}(x,s) + \frac{1}{p}\sigma_k \|s\|_2^{p+1}$$

Computation of the step:

1. minimize $m(s)$ until an approximate first-order minimizer is obtained:

$$\|\nabla_s m(s)\| \leq \kappa_{\text{stop}} \|s\|^p$$

Note: no global optimization involved.

Adaptive Regularization with p-th degree model

Algorithm 1.1: The ARp Algorithm

Step 0: Initialization: x_0 and $\sigma_0 > 0$ given. Set $k = 0$

Step 1: Termination: If $\|g_k\| \leq \epsilon$, terminate.

Step 2: Step computation:

Compute s_k such that $m_k(s_k) \leq m_k(0)$ and $\|\nabla_s m(s_k)\| \leq \kappa_{\text{stop}} \|s_k\|^p$.

Step 3: Step acceptance:

Compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - T_{f,p}(x_k, s_k)}$

and set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$

Step 4: Update the regularization parameter:

$$\sigma_{k+1} \in \begin{cases} [\sigma_{\min}, \sigma_k] & = \frac{1}{2}\sigma_k \quad \text{if } \rho_k > 0.9 & \textit{very successful} \\ [\sigma_k, \gamma_1 \sigma_k] & = \sigma_k \quad \text{if } 0.1 \leq \rho_k \leq 0.9 & \textit{successful} \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & = 2\sigma_k \quad \text{otherwise} & \textit{unsuccessful} \end{cases}$$

A few questions

$$f(x+s) \leq m(s) \stackrel{\text{def}}{=} T_{f,p}(x,s) + \tfrac{1}{p} L \|s\|_2^{p+1}$$

Obvious questions:

- can we avoid the global Lipschitz requirement? YES!
- can we approximately minimize m and retain good worst-case function-evaluation complexity? YES !
- does this work well in practice? yes for $p = 2$, unknown otherwise
- is it realistic to use high-degree derivative tensors? PROBABLY, for small dimensions, or partially separable problems

Derivative tensors for partially separable problems

f is partially separable if

$$f(x) = \sum_{i=1}^{m} f_i(U_i x) = \sum_{i=1}^{m} f_i(x_i) \quad \text{where} \quad \operatorname{rank}(U_i) \ll n$$

Then

$$\nabla_x^p f(x)[s]^p = \sum_{i=1}^{m} \nabla_{x_i}^p f_i(x)[U_i x]^p$$

Note:

$$\text{size}(\ \nabla_{x_i}^p f_i(x)\ ) \ll \text{size}(\ \nabla_x^p f(x)\ )\text{!!!}$$

Evaluation complexity: a challenging result

How many function evaluations (iterations) are needed to ensure that

$$\|g_k\| \leq \epsilon?$$

If ∇_x^p is globally Lipschitz, the ARp algorithm requires at most

$$\left\lceil \kappa_S \epsilon^{-\frac{p+1}{p}} \right\rceil \text{ evaluations}$$

for some κ_S independent of ϵ.

cf. Birgin, Gardenghi, Martinez, Santos, T. (MPA, 2017)

$\Longrightarrow$ increasing model degree improves complexity!

Evaluation complexity: proof (1)

$$f(x_k + s_k) \leq T_{f,p}(x_k, s_k) + \frac{L}{p}\|s_k\|^{p+1}$$

$$\|g(x_k + s_k) - \nabla_s T_{f,p}(x_k, s_k)\| \leq L\|s_k\|^p$$

Lipschitz continuity of $\nabla_x^p f(x)$

$$\forall k \geq 0 \qquad f(x_k) - T_{f,p}(x_k, s_k) \geq \frac{\sigma_{\min}}{p}\sigma_{\min}\|s_k\|^{p+1}$$

$$f(x_k) = m_k(0) \geq m_k(s_k) = T_{f,p}(x_k, s_k) + \frac{\sigma_k}{p}\|s_k\|^{p+1}$$

Evaluation complexity: proof (2)

$$\exists \sigma_{\max} \quad \forall k \geq 0 \qquad \sigma_k \leq \sigma_{\max}$$

Assume that $\sigma_k \geq \dfrac{L}{(1-\eta_2)}$. Then

$$|\rho_k - 1| \leq \frac{|f(x_k + s_k) - T_{f,p}(x_k, s_k)|}{|T_{f,p}(x_k, 0) - T_{f,p}(x_k, s_k)|} \leq \frac{Lp\|s_k\|^{p+1}}{p\,\sigma_k\|s_k\|^{p+1}} \leq 1 - \eta_2$$

and thus $\rho_k \geq \eta_2$ and $\sigma_{k+1} \leq \sigma_k$.

Evaluation complexity: proof (3)

$$\forall k \text{ successful} \qquad \|s_k\| \geq \left(\frac{\|g(x_{k+1})\|}{L + \kappa_{\text{stop}} + \sigma_{\max}} \right)^{\frac{1}{p}}$$

$$\begin{aligned}
\|g(x_k + s_k)\| &\leq \|g(x_k + s_k) - \nabla_s T_{f,p}(x_k, s_k)\| \\
&\quad + \left\| \nabla_s T_{f,p}(x_k, s_k) + \sigma_k \|s_k\|^{p-1} s_k \right\| + \sigma_k \|s_k\|^p \\
&\leq L_f \|s_k\|^2 + \|\nabla_s m(s_k)\| + \sigma_k \|s_k\|^p \\
&\leq [L_f + \kappa_{\text{stop}} + \sigma_k] \|s_k\|^p
\end{aligned}$$

Evaluation complexity: proof (4)

$\|g(x_{k+1})\| \leq \epsilon$ after at most $\dfrac{f(x_0) - f_{\text{low}}}{\kappa} \epsilon^{-(p+1)/p}$ successful iterations

Let $\mathcal{S}_k = \{j \leq k \geq 0 \mid \text{iteration } j \text{ is successful}\}$.

$$
\begin{aligned}
f(x_0) - f_{\text{low}} &\geq f(x_0) - f(x_{k+1}) \geq \sum_{i \in \mathcal{S}_k} \Big[f(x_i) - f(x_i + s_i) \Big] \\
&\geq \tfrac{1}{10} \sum_{i \in \mathcal{S}_k} \Big[f(x_i) - T_{f,p}(x_i, s_i) \Big] \geq |\mathcal{S}_k| \frac{\sigma_{\min}}{10p} \min_i \|s_i\|^{p+1} \\
&\geq |\mathcal{S}_k| \frac{\sigma_{\min}}{10p\Big(L_f + \kappa_{\text{stop}} + \sigma_{\max}\Big)^{\frac{p+1}{p}}} \min_i \|g(x_{i+1})\|^{\frac{p+1}{p}} \\
&\geq |\mathcal{S}_k| \frac{\sigma_{\min}}{10p\Big(L_f + \kappa_{\text{stop}} + \sigma_{\max}\Big)^{\frac{p+1}{p}}} \epsilon^{\frac{p+1}{p}}
\end{aligned}
$$

Evaluation complexity: proof (5)

$$k \leq \kappa_u |\mathcal{S}_k|, \quad \text{where} \quad \kappa_u \stackrel{\text{def}}{=} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right),$$

$\sigma_k \in [\sigma_{\min}, \sigma_{\max}]$ + mechanism of the σ_k update.

$\|g(x_{k+1})\| \leq \epsilon$ after at most $\dfrac{f(x_0) - f_{\text{low}}}{\kappa}\ \epsilon^{-\frac{p+1}{p}}$ successful iterations

One evaluation per iteration (successful or unsuccessful).

Evaluation complexity: sharpness

Is the bound sharp? YES for $p = 2$!!!

Construct a unidimensional example with

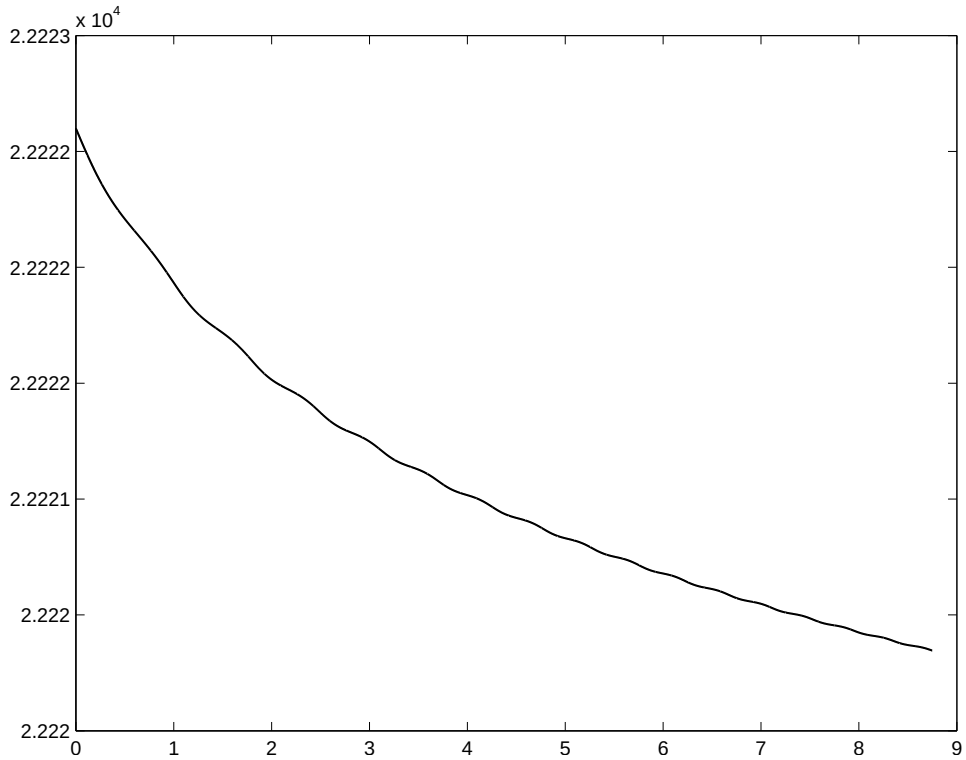
$$x_0 = 0, \quad x_{k+1} = x_k + \left(\frac{1}{k+1}\right)^{\frac{1}{3}+\eta},$$

$$f_0 = \frac{2}{3}\zeta(1+3\eta), \quad f_{k+1} = f_k - \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta},$$

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta}, \quad H_k = 0 \text{ and } \sigma_k = 1,$$

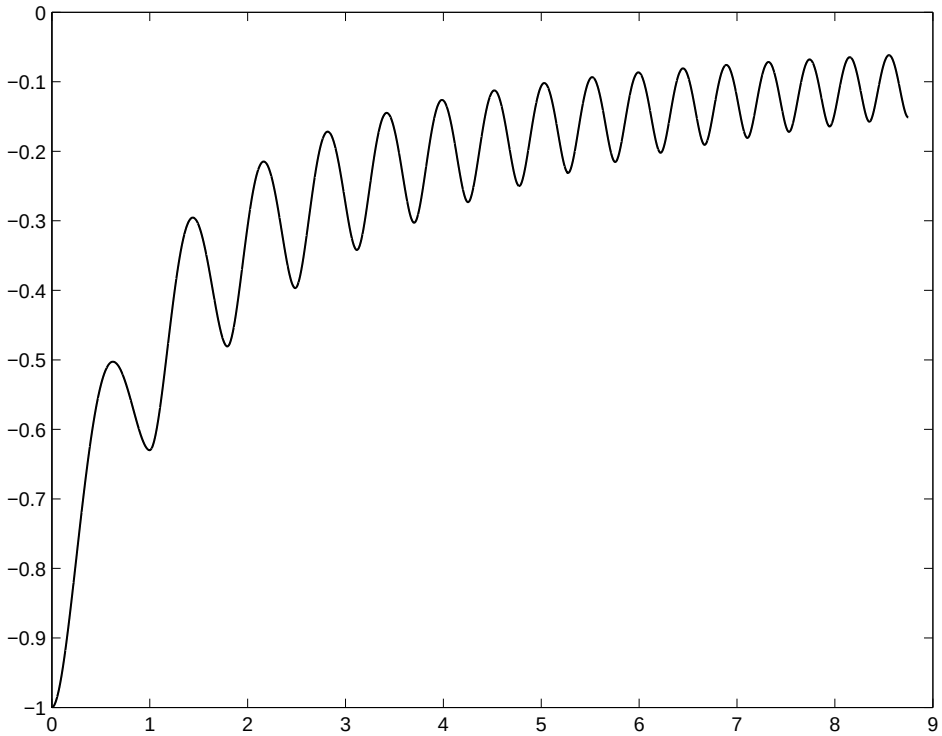
Use Hermite interpolation on $[x_K, x_{k+1}]$.

An example of slow ARC2 (1)



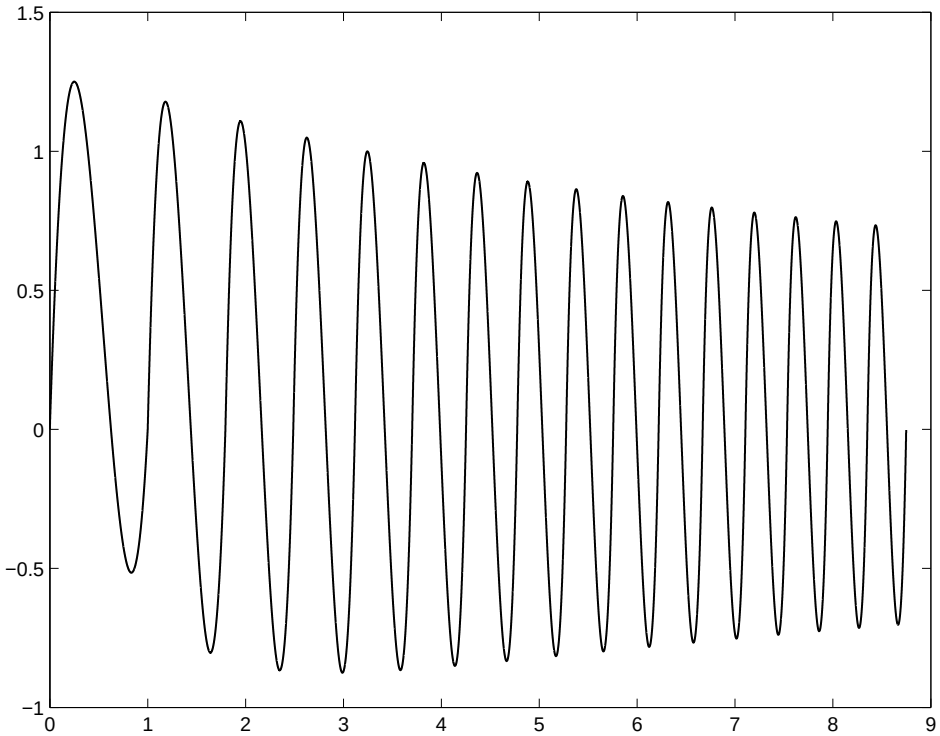
The objective function

An example of slow ARC2 (2)



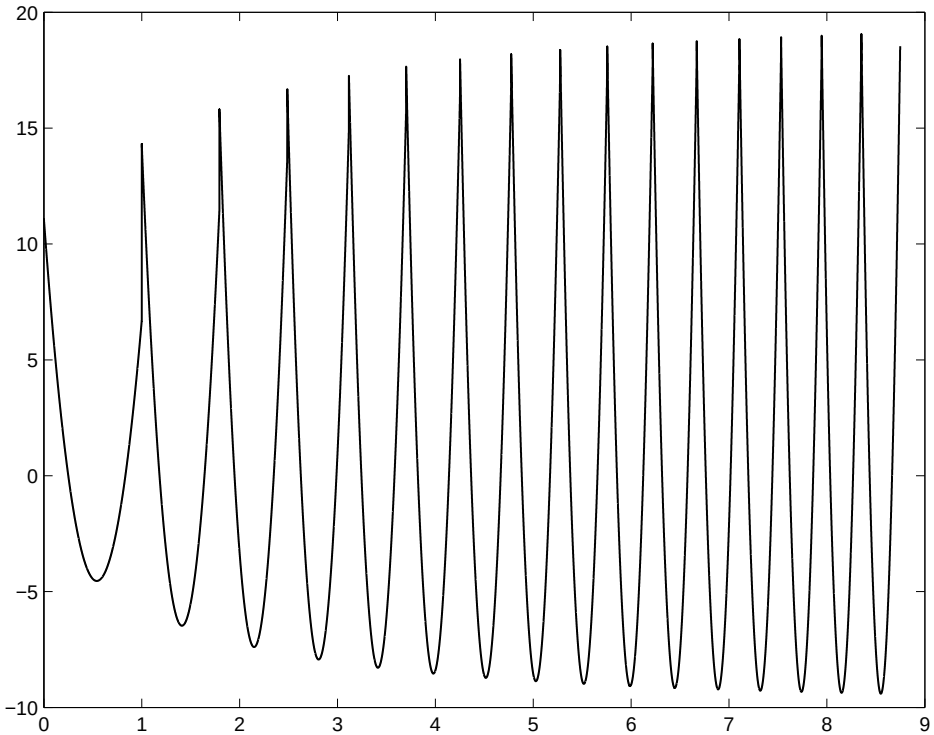
The first derivative

An example of slow ARC2 (3)



The second derivative

An example of slow ARC2 (4)



The third derivative

Standard methods are not so good

The steepest descent method with requires at most

$$\left\lceil \frac{\kappa_C}{\epsilon^2} \right\rceil \text{ evaluations}$$

for obtaining $\|g_k\| \leq \epsilon$.

Nesterov

Sharp??? YES

Newton's method (when convergent) requires at most

$$O(\epsilon^{-2}) \text{ evaluations}$$

for obtaining $\|g_k\| \leq \epsilon$

Sharp??? YES!!!

A (not so) obvious question

If one uses a model of degree p ($T_{f,p}(x,s)$), why be satisfied with first- or second-order critical points???

What do we mean by critical points of order larger than 2 ???

What are necessary optimality conditions for order larger than 2 ???

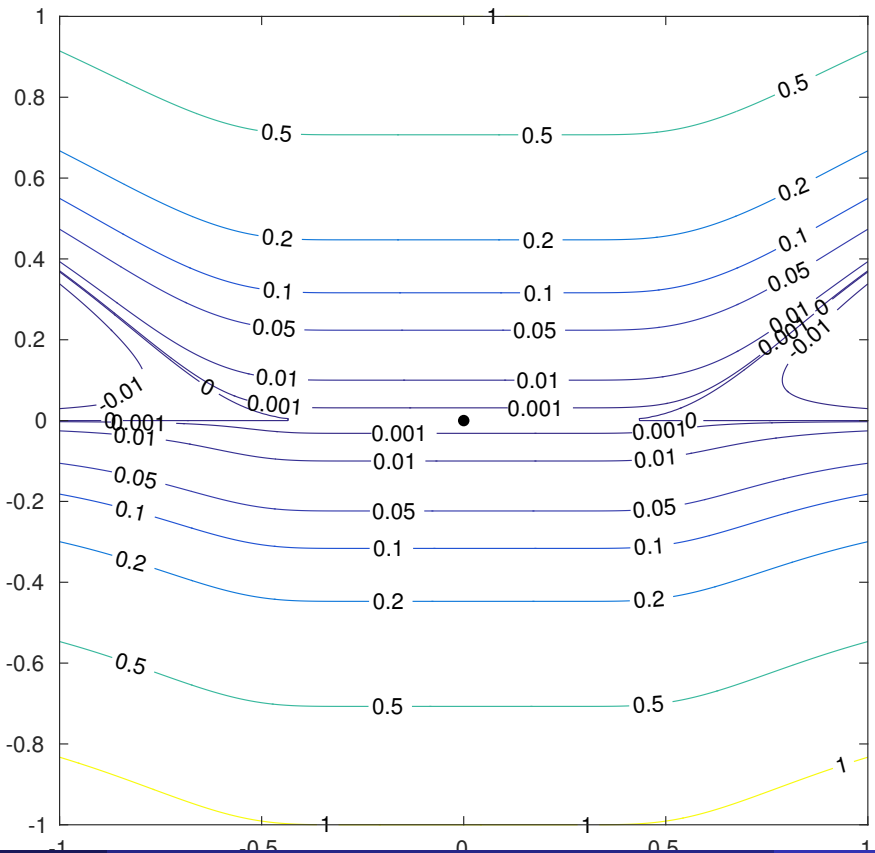
Not an obvious question!

A sobering example (1)

Consider the unconstrained minimization of

$$f(x_1, x_2) = \begin{cases} x_2 \left(x_2 - e^{-1/x_1^2} \right) & \text{if } x_1 \neq 0, \\ x_2^2 & \text{if } x_1 = 0, \end{cases}$$

(cf. Peano (1814), Hancock (1917))



A sobering example (2)

Conclusions:

- looking at optimality along straight lines is not enough
- depending on Taylor's expansion for necessary conditions is not always possible

Even worse:

$$f(x_1, x_2) = \begin{cases} x_2 \left(x_2 - \sin(1/x_1) e^{-1/x_1^2} \right) & \text{if } x_1 \neq 0, \\ x_2^2 & \text{if } x_1 = 0, \end{cases}$$

(no continuous descent path from 0, although not a local minimizer!!!)

Hopeless?

Limiting one's ambitions...

Note: the non-existence of continuous descent paths remains a necessary condition! Focus on polynomial paths

$$x(\alpha) = x_* + \sum_{i=1}^{q} \alpha^i s_i + o(\alpha^q)$$

Suppose that x_* is a local minimizer. Then, for $j \in \{1, \ldots, q\}$,

$$\sum_{k=1}^{j} \frac{1}{k!} \Big(\sum_{(\ell_1,\ldots,\ell_k) \in \mathcal{P}(j,k)} \nabla_x^k f(x_*)[s_{\ell_1}, \ldots, s_{\ell_k}] \Big) \geq 0$$

holds for all $(s_1, \ldots, s_j)$ such that, for $i \in \{1, \ldots, j-1\}$,

$$\sum_{k=1}^{i} \frac{1}{k!} \Big(\sum_{(\ell_1,\ldots,\ell_k) \in \mathcal{P}(i,k)} \nabla_x^k f(x_*)[s_{\ell_1}, \ldots, s_{\ell_k}] \Big) = 0.$$

And then?

In short:

- reduces to (in)equalities on $\nabla_x^j f(x)[s]^j$ in the kernel of $\nabla_x^j f(x)[s]^{j-1}$ for j = 1,2,3
- inherently more complicated for orders 4 and above (conditions involving a mix of $\nabla_x^j f(x)[s]^j$ of different orders)

Desperate?

Using Taylor's models, nevertheless

Define, for some small $\Delta > 0$,

$$\phi^{\Delta}_{f,j}(x) \stackrel{\text{def}}{=} f(x) - \underset{\substack{x+d\in\mathcal{F}\\ \|d\|\leq\Delta}}{\text{globmin}}\ T_{f,j}(x,d),$$

$$\left[\lim_{\Delta\to 0} \frac{\phi^{\Delta}_{f,j}(x)}{\Delta^j} = 0\right] \Rightarrow \quad \text{path-based necessary conditions at } x \quad .$$

$\nabla^q_x f$ Lipschitz continous near $x_\epsilon \in \mathcal{F}$. Suppose that

$$\phi^{\Delta}_{f,j}(x_\epsilon) \leq \epsilon\Delta^j \quad \text{for} \quad j = 1,\ldots,q$$

Then

$$f(x_\epsilon + d) \geq f(x_\epsilon) - 2\epsilon\Delta^q \quad \forall x_\epsilon + d \ \text{ with } \ \|d\| \leq \left(\frac{q!\,\epsilon\Delta^q}{L_{f,q}}\right)^{\frac{1}{q+1}}$$

A (theoretical) trust-region algorithm

Algorithm 2.1: Trust-region with adaptive order models (TRq)

Step 0: Initialization: q, $\epsilon \in (0,1]$, x_0 and $\Delta_1 \in [\epsilon, 1]$, $\Delta_{\max} \in [\Delta_1, 1]$.

Step 1: Step computation: For $j = 1, \ldots, q$,
(i) evaluate $\nabla^j f(x_k)$ and $\phi_{f,j}^{\Delta_k}(x_k)$
(ii) if $\phi_{f,j}^{\Delta_k}(x_k) > \epsilon \Delta_k^j$, go to Step 3 with $s_k = d$
Terminate with $x_\epsilon = x_k$ and $\Delta_\epsilon = \Delta_k$.

Step 2: Accept the new iterate: Compute $f(x_k + s_k)$ and

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_{f,j}(x_k, 0) - T_{f,j}(x_k, s_k)}.$$

If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise set $x_{k+1} = x_k$.

Step 4: Update the trust-region radius. Set

$$\Delta_{k+1} \in \begin{cases} [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k < \eta_1, \\ [\gamma_2 \Delta_k, \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\Delta_k, \min(\Delta_{\max}, \gamma_3 \Delta_k)] & \text{if } \rho_k \geq \eta_2, \end{cases}$$

An evaluation complexity bound

TRq computes an ϵ-approx "q-th order critical point" after at most

$$\kappa_S \, \epsilon^{-(q+1)}$$

(successful) iterations, and the bound is essentially sharp.

Same results for problems involving convex constraints!

[needs a "constraint qualification" for the boundary of the feasible set]

Conclusions

Evaluation complexity improves with the model's degree

Critical points of order higher than 2 are (in general) evasive

Approximate critical points high order can be defined

An evaluation complexity bound for those is available
(more work for higher orders)

The above holds for unconstrained and convexly constrained problems

Further questions

Can one improve the complexity bound for $p > q$???

What about equality constrained problems?

Can this be (more) practical?

Many thanks for your attention!

Reference:

C. Cartis, N. Gould and Ph. L. Toint,

"Second-order optimality and beyond: characterization and evaluation complexity in convexly-constrained nonlinear optimization", FoCM, to appear.